AI AGENTS AND MORAL RESPONSIBILITY

Roger Crisp (Oxford University)

Artificial agents – such as LLMs – are already 'doing' things, such as composing new poems, or helping students with their homework. But we tend to use the notions of 'doing' or 'acting' very widely, allowing a stone, for example, to break a window, or the wind to carry me off course in my journey.

A stone, or the wind, however, does not have any wishes, desires, or intentions, while we might speak of an LLM being in such a state, perhaps so as to explain how some particular process has led to some particular result or other. But such claims might be said to be largely metaphorical. And even if they are not, such states are still not sufficient to hold their possessor morally responsible. The LLM is doing what it does purely because of the material state underlying it, a state which has been designed by human beings. If any agent is morally responsible for some state of affairs produced by an LLM, then, won't it be the human being who designed it?

But still we have to say more: what if that human being was themselves hypnotized and then commanded to create the LLM? Their action would not be fully free, though that of the hypnotizer is. So for moral responsibility what really matters is freedom.

I now want to suggest that, as far as freedom is concerned, we humans are in the same condition as AI agents. True, we are not ourselves 'artificial', the product of art, skill, intelligence, or design. But that does not matter. Like AI agents, we are also the results of natural processes and events over which we have no control, and hence for which we have only the same degree of responsibility as AI agents.

What do I mean here by 'responsibility'? I am not talking about *causal* responsibility, as in: 'The wind was responsible for my ending up in Aegina rather than Athens'. Nor am I speaking of *legal* responsibility, as in: 'The driver of a car is *legally* responsible for any damage caused by their driving'. Rather, *moral* responsibility concerns that for which we are *morally* accountable, that for which we can be blamed, or praised, morally.

What are the conditions for such responsibility? One key condition is that the action must be 'up to us', as it will not be in cases like that of hypnotism. But it is hard to see how any action can meet this condition.

Consider first what Peter van Inwagen calls the 'consequence argument': 'If determinism is true, then our acts are the consequences of the laws of nature and events in the remote past. But it is not up to us what went on before we were born [i.e., we do not have the ability to change the past], and neither is it up to us what the laws of nature are [i.e., we do not have the ability to break the laws of nature]. Therefore, the consequences of these things (including our present acts) are not up to us' (van Inwagen).

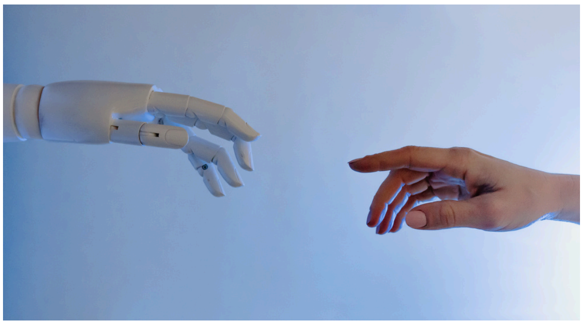
A second argument is Galen Strawson’s ‘basic argument’: ‘(1) Nothing can be *causa sui* - nothing can be the cause of itself. (2) In order to be truly morally responsible for one's actions one would have to be *causa sui*, at least in certain crucial mental respects. (3) Therefore nothing can be truly morally responsible’.

Consider now another argument related to both of the above, which we might call the ‘belief-desire’ argument: (1) Any action is fully explicable with references to the beliefs and desires of the agent that caused it. (2) An agent’s beliefs and desires are not themselves willed by that agent. (3) Hence an agent cannot be held responsible for the beliefs and desires that led to their action, and hence for that action itself.

Finally, let me mention the ‘Manipulation Argument’, stated by Al Mele as follows: ‘Diana creates a zygote Z in Mary. She combines Z’s atoms as she does because she wants a certain event *E* to occur thirty years later. From her knowledge of the state of the universe just prior to her creating Z and the laws of nature of her deterministic universe, she deduces that a zygote with precisely Z’s constitution located in Mary will develop into an ideally self-controlled agent who, in thirty years, will judge, on the basis of rational deliberation, that it is best to A and will A on the basis of that judgment, thereby bringing about *E’*. It seems quite unfair and unreasonable to hold Mary morally responsible for bringing about *E*. And it seems each of us is always in the same position as Mary, because our actions are the result of natural laws over which we have no control.

Am I claiming, then, that just as AI agents cannot plausibly be held morally responsible, neither can human agents like us?

Yes, if we insist that true freedom is a necessary condition for holding agents morally responsible. But there is another way to justify practices of holding responsible: considering the consequences of those very practices. It is in one respect ‘unfair’ to hold any agent morally responsible; but such practices may themselves have good consequences, through, for example, deterring the agent from acting in the same way in future. If we understand blaming as a form of ‘punishment’, that is to say, we must understand its justification not as ‘retributive’ (since no agent ever ‘deserves’ to be harmed in any way), but as ‘forward-looking’ or ‘consequentialist’. Our practices of holding responsible, if they are well designed, will make the world better than it would otherwise have been; and if artificial agents can themselves respond to being blamed in the same way as we are inclined to respond, they can be blamed (and of course praised) just as we are and for the same reasons.



Roger Crisp

Director, Oxford Uehiro Centre for Practical Ethics
Professor of Moral Philosophy, University of Oxford
Uehiro Fellow and Tutor in Philosophy, St Anne’s College, Oxford